ORIGINAL

NATIONAL UNIVERSITY OF SINGAPORE

EE5907– PATTERN RECOGNITION

(Semester 1: AY2018/2019)

Time Allowed: 2.5 Hours

INSTRUCTIONS TO STUDENTS

1. Please write only your Student Number. Do not write your name.
2. This assessment paper contains **FOUR (4)** questions and comprises **SIX (6)** printed pages.
3. Students are required to answer **ALL** questions.
4. Students should write the answers for each question on a new page.
5. This is a **CLOSED BOOK** assessment. One A4-size formula sheet is allowed.
6. Non-programmable calculators are allowed.
7. Total Marks is **ONE HUNDRED (100)**.

Q1 (25 marks). Subquestions (a) and (b) can be answered independently.

(a) Consider a 2-class naive Bayes classifier with one binary feature and one Gaussian feature. More specifically, class label y follows a categorical distribution parametrized by π, i.e., $p(y = c) = \pi_c$. The first feature x_1 is binary and follows a Bernoulli distribution: $p(x_1|y = c) = \text{Bernoulli}(x_1|\theta_c)$. The second feature x_2 is univariate Gaussian: $p(x_2|y = c) = \mathcal{N}(x_2|\mu_c, \sigma_c^2)$. Let $\pi = [0.8 \;\; 0.2]$, $\theta = [0.2 \;\; 0.7]$, $\mu = [-1 \;\; 1]$ and $\sigma^2 = [1 \;\; 1]$.

(i) Compute $p(y|x_2 = 1)$. Note that result is a vector of length 2 that sums to 1.

(7 marks)

(ii) Compute $p(y|x_1 = 0, x_2 = 1)$.. Note that result is a vector of length 2 that sums to 1.

(6 marks)

(b) Consider a binary classification problem of predicting binary class y from features x. The cost of correct prediction is \$0. There is a \$5 cost associated with predicting class 0 when the true class is 1. There is a \$2 cost associated with predicting class 1 when the true class is 0. Suppose the cost of asking a human to perform the manual classification is \$1. Therefore, for a particular x, there are three possible decisions: (1) decision α_0 predicts y to be 0, (2) decision α_1 predicts y to be 1 and (3) decision α_h requires a human to perform the manual classification. Let $p_1 = p(y = 1|x)$

(i) Assume the human is 100% accurate. What is the general decision rule (as a function of p_1) in order to minimize expected loss?

(6 marks)

(ii) Assume the human is only 90% accurate. Assume that when the human is wrong, the correct class is equally likely to be class 0 or class 1. What is the general decision rule (as a function of p_1) in order to minimize expected loss?

(6 marks)

Q2 (25 marks). Subquestions (a), (b) and (c) can be answered independently.

(a) Consider a geometric distribution $p(X = k) = (1 - q)^{k-1}q$, where k is the number of coin tosses until a head appears and q is the probability of getting a head for a particular coin toss. Suppose we observe N independent samples from the geometric distribution: $D = \{x_1, \cdots, x_N\}$.

(i) What is the maximum likelihood (ML) estimate of q? Please show your steps to get full credit.

(9 marks)

(ii) Suppose we use ML estimate of q to predict new data x_{N+1}. What problems might arise? Describe a solution to avoid this problem.

(2 marks)

(b) Consider the same distribution and data from part (a). It turns out that the conjugate prior distribution of the geometric distribution is the beta distribution: $p(q) = \frac{1}{B(a,b)} q^{a-1}(1 - q)^{b-1}$, where $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

(i) The posterior distribution $p(q|D)$ is a Beta distribution with parameters a', b'. What are a' and b'? Show your steps.

(6 marks)

(ii) What is the posterior predictive distribution $p(x_{N+1}|D)$? Show your steps. Your final answer can contain B.

(8 marks)

Q3 (25 marks). Given labelled -dimensional training vectors $x \in R^d$ from C classes, with $n_i$vectors from class c_i for $i = 1, \ldots, C$ and $\sum_{i=1}^{C} n_i = n$. Linear discriminative analysis (LDA) projection direction, $W \in R^{d \times p}$, is obtained by maximizing

$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|}$$

where $S_B \in R^{d \times d}$ and $S_w \in R^{d \times d}$ are the between class scatter and within class scatter respectively.

(a) Derive the expression for the optimal projection direction W^* for $C = 2$. (10 marks)

(b) Fish caught at sea are all mixed together in the net and are separated at a factory for packaging. It is required to design a pattern recognition system to separate flounder from cod by taking an image of the fish on a conveyer belt and then classifying and sorting the fish for processing. Measurements of length and width for 10 fish of each kind are obtained to develop a training set and are given in the table below.

COD		FLOUNDER	
Length	Width	Length	Width
18.5	5.3	10.4	8.3
21.5	6.4	13.5	9.9
15.6	6.6	11.9	8.6
19.9	5.4	12.6	7.8
8.5	3.2	8.2	5.3
16.4	5.6	13.2	8.2
17.3	5.9	18.4	10.6
12.7	3.4	6.1	4.2
22.5	8.3	17.3	9.5
12.9	4.8	14.6	7.8
165.8	54.9	126.2	80.2

SUM

(i) The within-scatter matrices for cod and flounder, S_c and S_f, are as follows:

$$S_c = \begin{bmatrix} 169.40 & 48.79 \\ 48.79 & 20.27 \end{bmatrix}, \; S_f = \begin{bmatrix} 127.84 & 59.06 \\ 59.06 & 34.92 \end{bmatrix}$$

If it is desired to devise a classifier based on one feature only, which would you choose, length or width? Given reasons. What is the least error that can be achieved on the training data with the chosen feature? Write down the decision rule for the classifier that achieves this.

(10 marks)

(ii) If it is desired to classify based on only one feature, it might also appear reasonable to perform dimensionality reduction using PCA, and then classify in the reduced-dimension space. Comment for or against this idea. Suggest an alternative if you are against this idea.

(5 marks)

Q4 (25 marks). The following sub-problems are independent.

(a) Given four data points $x_1 = \binom{0}{0}, x_2 = \binom{2}{0}, x_3 = \binom{0}{2}, x_4 = \binom{2}{2}$, please apply k-means to partition the data points into two clusters using following two initial centroids respectively: (i) the initial centroids are $\{x_1, x_2\}$, please list the iterative centroids until convergence; (ii) the initial centroids are $\{x_2, x_4\}$, please list the iterative centroids until convergence.

(5 marks)

(b) The sensitivity to initial centroids is a key issue for k-means, please list the popular solutions to selection on proper centroids.

(5 marks)

(c) What are the general differences between generative models and discriminative models? Given three example algorithms for each type of models.

(5 marks)

(d) Follow Kuhn-Tucker theorem to convert the primal constrained optimization problem in two class Support Vector Machine (SVM) into a dual, unconstrained one. The primal optimization problem is given by:

$$\begin{aligned} &\min \ \frac{1}{2}\|w\|_2^2 \\ &\text{s.t. } y_i w^T x_i \geq 1, i = 1, \ldots, n \end{aligned}$$

where (x_i, y_i) is a training data, $x_i \in R^d$ is a feature and $y_i \in \{-1, +1\}$ is the corresponding label. Derive the following dual form:

$$\max \ -\frac{1}{2}\sum_{j,k=1}^{n} \alpha_j \alpha_k y_j y_k (x_j^T x_k) + \sum_{j=1}^{n} \alpha_j.$$

(10 marks)

END OF PAPER